Breakout Session 6:

Migration to Cloud of the Oncogenomics Next Generation Sequencing Pipelines & Databases for CCDI and Other Pediatric Cancers

> Dr. Javed Khan Senior Investigator, CCR, NCI

Migration to Cloud of the Oncogenomics Next Generation Sequencing Pipelines & Databases for CCDI and Other Pediatric Cancers

PI: Dr. Javed Khan

Current Project Personnel: Hsien-Chao Chou, Vineela Gangalapudi, Vishal Koparde, Jun Wei and Patrick Zhao

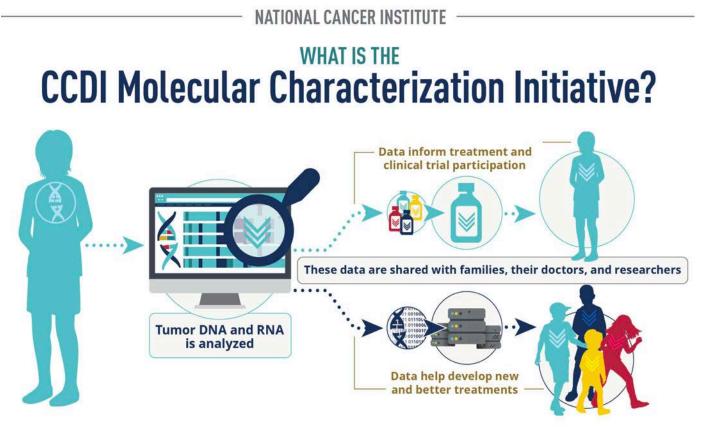
AWS ProSupport: Kevin Sayers

2024 NIH/ODSS Cloud Supplement Program PI Meeting January 17-18, 2024

Overarching Goals

- Develop cloud-based state-of-the art next generation sequencing analysis (NGS) pipeline based on best practice
- Develop a secure cloud-based comprehensive, genomic integrated browser for visualization and exploration of clinical (patient-centric) and research (cohort summary data) for pediatric cancers
- Support hypothesis generation, publications, and grant applications
- Complimentary to existing browsers (cBioPortal, Genomic Data Commons Data Portal, St. Jude Cloud, and UCSC Treehouse)
- Building on the existing Oncogenomics Clinical NGS pipeline and ClinOmics Portal

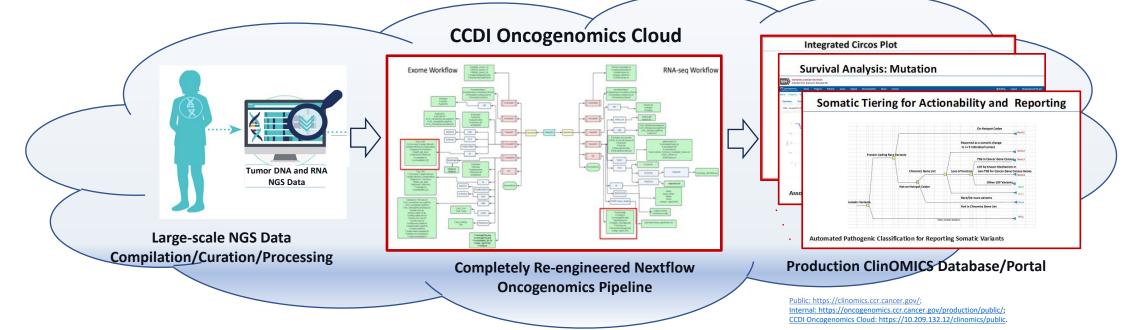
Background: Childhood Cancer Data Initiative (CCDI) Molecular Characterization Initiative (MCI)



Source: https://www.cancer.gov/research/areas/childhood/childhood-cancer-data-initiative

Generate clinical grade next generation sequencing (NGS) and methylation data from every child, adolescent, and young adult diagnosed with childhood cancer, enrolled in Children's Oncology Group (COG) trials nationwide

Childhood Cancer Data Initiative (CCDI) Oncogenomics Cloud Project



<u>Data</u>

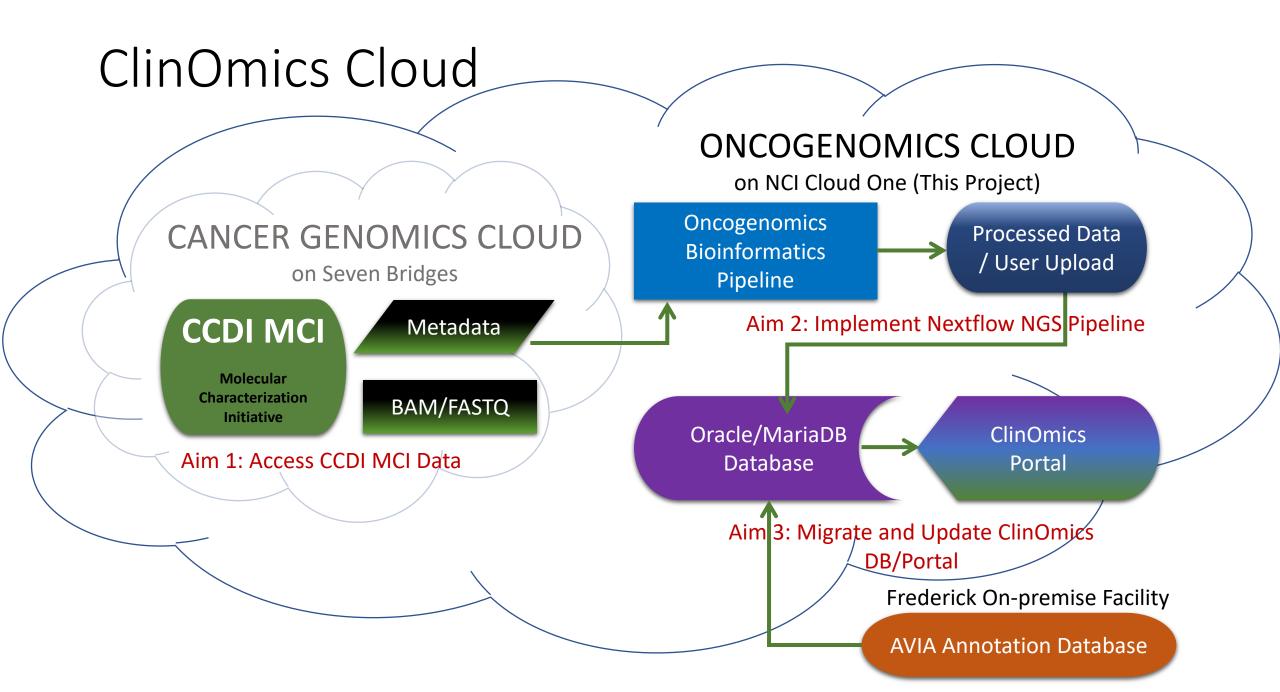
Large-scale tumor Whole Transcriptome and tumor/normal Whole Exome <u>from every child and</u> <u>adolescent young adult</u>

Oncogenomics NGS Pipeline

Processing and analyzing largescale Transcriptomic and WES datasets

ClinOmics DB/Portal

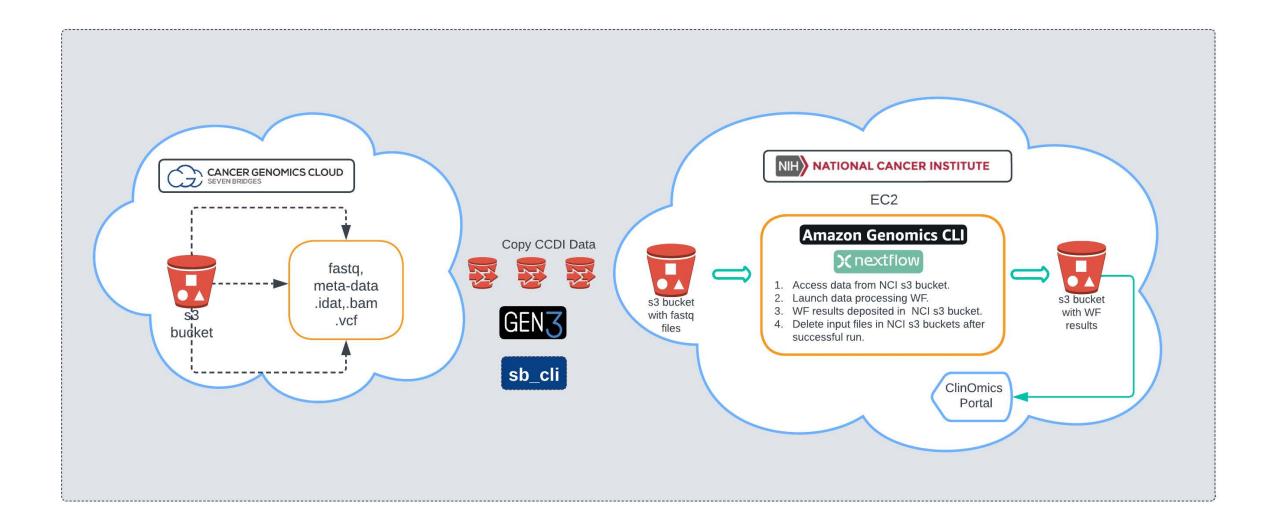
Bring the CCDI MCI and other pediatric cancer genomics data to the fingertips of researchers and clinicians.



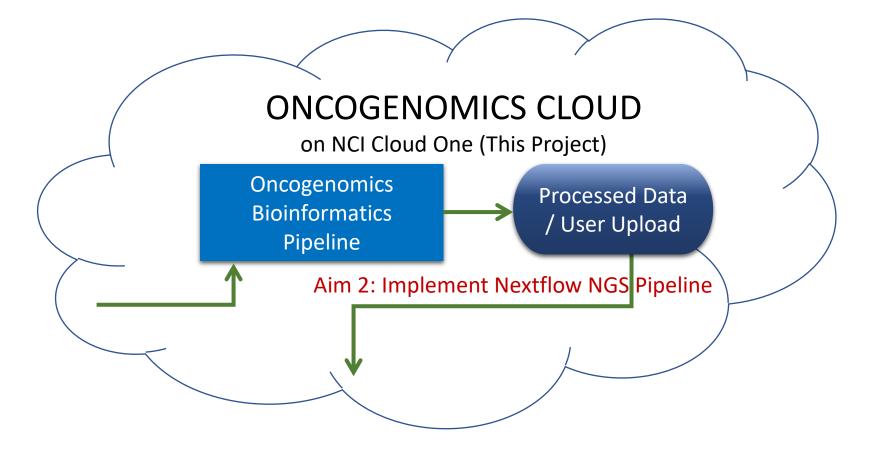
Aim #1: Directly accessing the CCDI Molecular Characterization Initiative (MCI) data in Cancer Genomics Cloud (CGC)

- Explore security and data sharing policies and technological means to directly access the CCDI MCI data in CGC hosted in Seven Bridges.
- Direct access to MCI data in CGC would avoid or reduce data download/egress costs of data analysis via S3 bucket sharing in the AWS cloud.
- The cross-cloud architecture would improve the data transfer, computational, and cost efficiencies of research.
- Pathway to access other childhood cancer genomics data hosted in AWS cloud

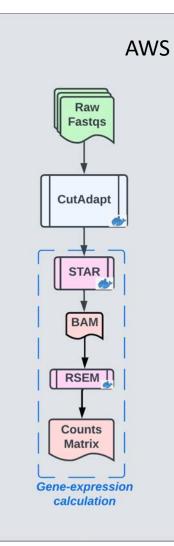
Aim #1 Ongoing: Accessing MCI Data in CGC



Aim #2: Develop and Deploy Oncogenomics NGS Pipeline in the AWS Cloud



Oncogenomics Pipeline – Conversion to Nextflow



	2	Crocon
Technical Feature/Metric	Snakemake	X nextflow
dryrun	\checkmark	×
"bin" folder	×	\checkmark
"variable" input	×	\checkmark
HPC<->Cloud interoperability	×	\checkmark
s3 fs support	×	\checkmark
direct GitHub support	×	\checkmark
inbuilt SLURM support	×	\checkmark
better regex support	×	\checkmark
code simplicity	\checkmark	×

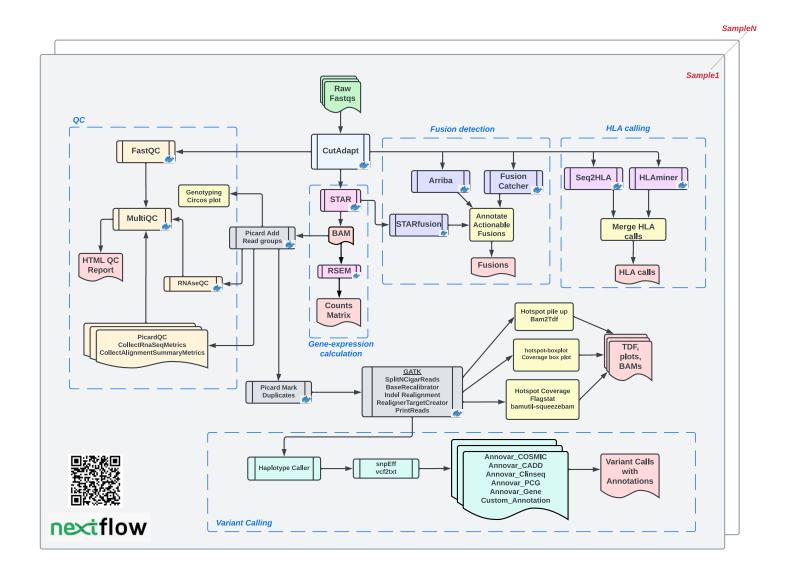
Achieved: Minimal Viable Product (MVP) in Year 1

Re-engineered Oncogenomics RNA-seq Pipeline, deployable on Biowulf, AWS Cloud Computing Platform, and other High-performance environments.

RNASeq

- Gene Expression
- QC
- Fusion Detection
- HLA Calling
- Variant Calling
- HLA

Easy to append new features or switch between genome versions (hg19<->hg38)



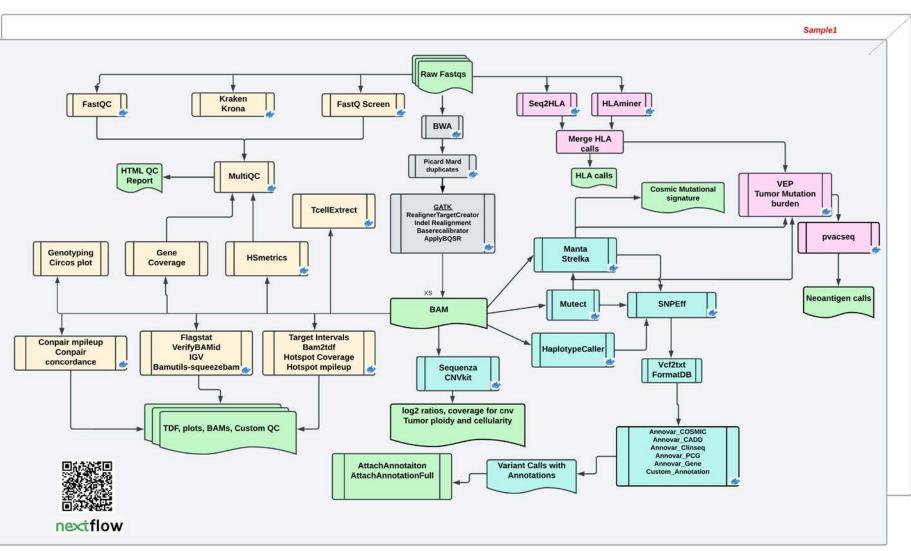
Ongoing: Exome Workflow – Year 1, 2

Re-engineered Oncogenomics Exome Pipeline, deployable on Biowulf, AWS Cloud Computing Platform, and other High-performance environments.

Data generated

- Actionable Germline
- Actionable Somatic
- HLA
- Copy number variant
- Mutation signature
- Mutational burden
- MSI
- Neoantigen
- Immune infiltrate

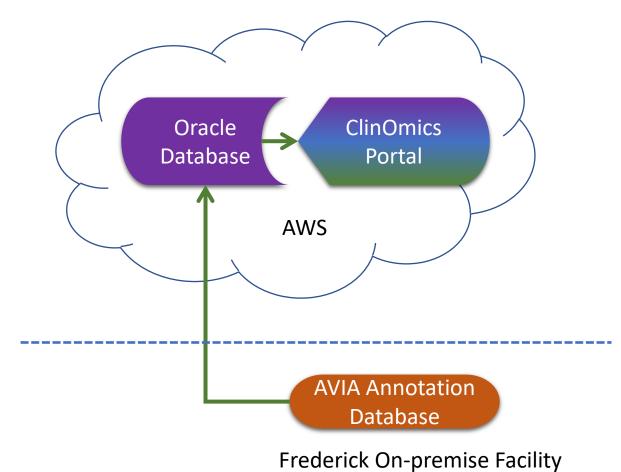
Easy to append new features or switch between genome versions (hg19<->hg38)



SampleN

Aim #3: Migrating DB/Web Portal to AWS

https://clinomics.ccr.cancer.gov/



Cloud/On-premise Databases Sync for Up-to-date AVIA Annotations

Oncogenomics DB for Clinical and Research Applications

<u>Public: https://clinomics.ccr.cancer.gov/;</u> Internal: https://oncogenomics.ccr.cancer.gov/production/public/; Cloud (This Project): https://10.209.132.12/clinomics/public.

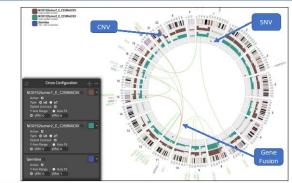
Multi-omics individual patient portal



	Lo	oll	lip	0	o Plo	ot				Mut: pi	atio oje		n		orted M n TCGA/	utations ICGC	
Reported amattores] 						ini.	URAS CB	-			-	-		136) Mutatio originarymaa orieshift detti orieshift detti originarishift de originarishift de originarishift or topgen 0(25)	SAV 7(1261) V 0(24) as 0(10) ios 0(3) Retion 0(11)	
								atein pos	Rivers.								
ALC: UNK	0 20	at to	open.	4 1109 128						also a	Carolina I	÷	(ingenit)	Patients	: 5/10 V	ariants:	7/
MAT:	н	n Total Co	er a	A new tak	0	• Tier:	1	5 4	na vier 🛛 Al	Hag Por	ierts (1	±	 Diage 	rosis: Any		ariants:	7/
MAT:			er a		0	Tier:	1	5 4	no tier Al	Hing Por	uns (tin Ini	 Diagram 				7/
tar:	н		er a		Dispose 8	Tier: Patient	t t	a a a	NUTER AL	Hap Por	uens (2	<u>ل</u>	• Diag	iosis: Any Searc	84 [•	7/
NAT: [] Show]	5 T entries	Series C	ev) (0 adames) sta vát		Patient g	1 7 Case 1 201000115	Gut	anne an	100.0	-		Game	iosis: [Any Searc	84 [•	
NAT: [] Show]	5 T estries	Select C	ine: 0 interest content	Silte colourt	Disposite	Pattern (al		aut aa	Start 140453136	100.0	8-11 A		Group	AAChungel	lintepotes	sup118 (C
NAT: [] Show]	5 T entries Liberation	Senat D	ivi (0 idatore contact 13.7%	Min VAF	Dispose	Patient MC30072 NC30037	20160415	art art	Stan 1 (140453)(36 (140453)(36	101651136	8. A		Group	AACDange VECOE	tiniquite!	* Septid (00
NAT: [] Show]	ST outries	Series D	ier (0 identes calcular 13.7%	Min VAF	Disposes (Nelsoyna Nelsoyna	Putlion (14 16(30)72 16(30)72	20160415 20161018	Out 0/7 0/7 0/7	Start 140453136 140453136 140453136	140453136	8-41 A A A		Group BRAF	AADampo Vices Vices		* Sept104 (0113080022 0113080022	000
MAT: [] Show]	ST outries Libraries Q Q Q	Series O	13.7% 13.7%	Min VAF Colort 6.85% 6.85% 6.85%	Disposi Nelesens Nelesens Relesens	Puttern () MC90872 MC90377 MC90277 MC90272 MC90291	20160415 20161018 20160929	0u1 0.7 0.7 0.7	5444 (140453136 140453136 140453136	149453136 149453138 149453138	A A A A		Grand MAR DUAR 20147	AADangel Vitor Vitor Vitor Vitor	* 0 0 0 0 0 0 0 0 0 0 0	* 6+++138 ++113880012 ++113480012 ++113480022	0000
MAT: [] Show]	ST outries Ultraction 0 0 0 0	Lener O	er 0 darms cabert 13.7% 13.7% 13.7% 13.7%	585 VAF colored 6.85% 6.85% 6.85% 6.85%	Disposis d Relation Malanaria Relations Oblighted	Putleon (MC30072 MC30077 MC30077 MC30077 MC30072 MC30072	20160415 20161018 20160929 20160415	041 0/7 0/7 0/7 0/7	Stan (140453136 140453136 140453136 140453136 140453136	End 1 169453136 149453138 149453138	a A A A A		Enter Enter Ester Ester	AADampel AADampel Victor Victor Victor Victor		* Sep138 (0113880022 0113880022 0113880022 0113880022	00000

Publications, grant applications, data exploration, biological insights

Integrated Circos Plot



Clinical reports, publications, grant applications

IGV Viewer across all samples for one patient

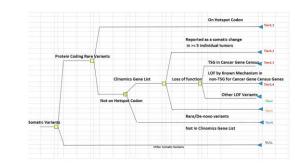


Visual Validation of Variant, Germline vs. Somatic, Tumor evolution

Survival Analysis: Mutation



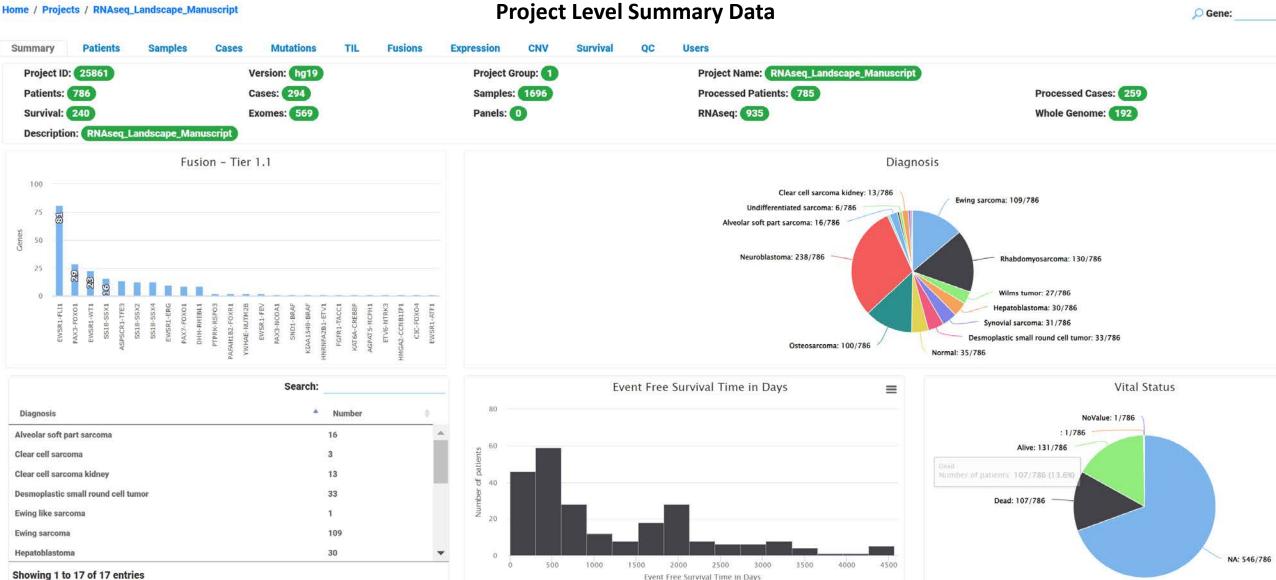
Somatic Tiering for Actionability and Reporting



Association of MYOD1 Mutation with Outcome in Rhabdomyosarcoma

Automated Pathogenic Classification for Reporting Somatic Variants

Oncogenomics DB for Clinical and Research Applications



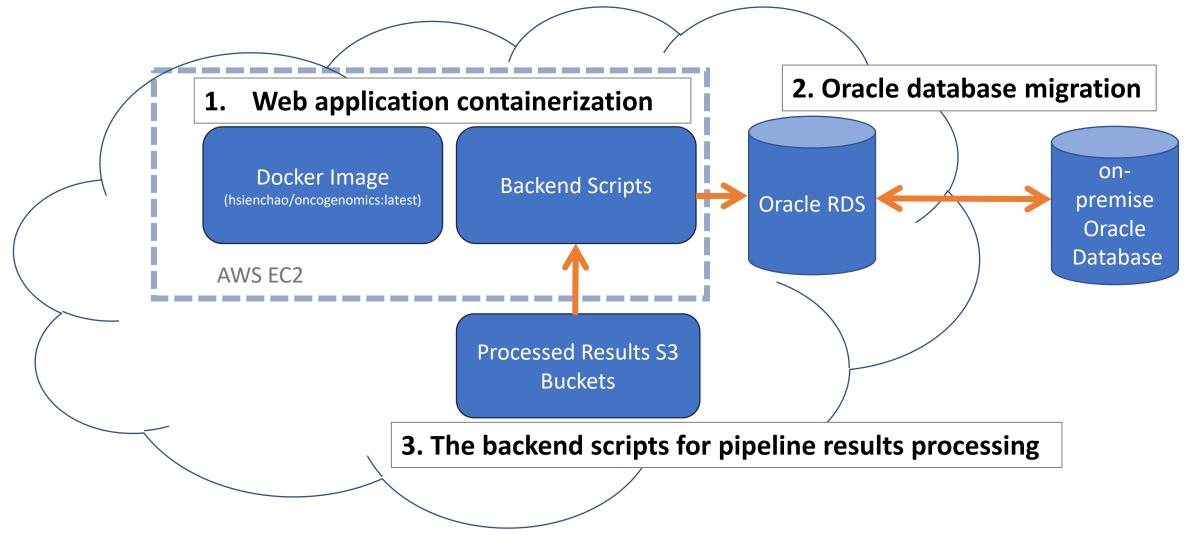
Variant annotation – AVIA (Annotation, Visualization and Impact Analysis)

- Seamless integration with AVIA (<u>https://avia-abcc.ncifcrf.gov/</u>)
- AVIA provides up-to-date customized OpenCRAVAT based annotation
 - Reported database (CBioPortal, ICGC, Genie and PCG)
 - Population frequency
 - Protein annotation
 - Functional prediction
 - Epigenetics information
 - Clinical information
 - Basic gene information

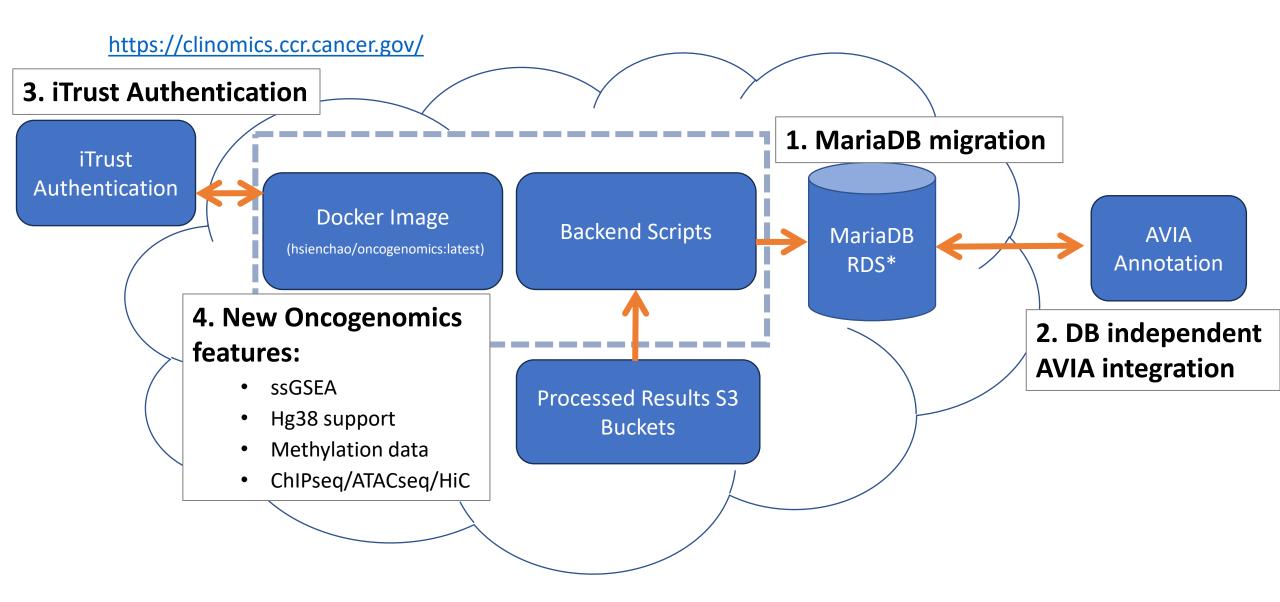
Gene: PTH2		Chr: chr19	Start: 499265	33 End	: 49926533	Ref: G	Alt: C
Population	Protein	Prediction	Epigenetics Re	ported Cl	inical Ger	ne	
ICGC		Annotato	or: ICGC				
Cbioportal		Descript	ion: ICGC				
Pediatric		Column	Description			Value	
		ICGC	ICGC	Total count: 16 ID: MU46003			
				Project	Donors	Total Donors	Frequency
				ESCA-CN	2	332	0.00602
				LICA-CN	9	402	0.02239
				LINC-JP	1	394	0.00254
				LICA-FR	1	252	0.00397
				GACA-JP	3	585	0.00513

Achieved: Migrating DB/Web Portal to AWS

https://clinomics.ccr.cancer.gov/



Ongoing: Migrating DB/Web Portal to AWS



Achievements and future work

achievements	Achievements/Products	Planned Future Work
Pipeline	Nextflow RNA- and DNA-seq Pipeline: https://github.com/CCRGeneticsBranch/AWS_POC_MVP_NF	 Add Whole Genome and Methylation workflow modules
Рір	Pipeline Module Docker Containers: https://github.com/CCRGeneticsBranch/Dockers	worknow modules
Web	The Oncogenomics Database and Web Portal: https://10.209.132.12/clinomics/public	
and tal	The Oncogenomics Portal Source Code: https://github.com/CCRGeneticsBranch/Oncogenomics_v2	 Incorporate methylation data and tools, including diagnostic classification (TSNE plots (
Database Por	The Oncogenomics Portal Docker: https://github.com/CCRGeneticsBranch/Oncogenomics_docker	classification/ TSNE plots/ deconvolution

Acknowledgements

Genetics Branch

- Patrick Zhao
- Vineela Gangalapudi
- Hsien-chao Chou
- Jun Wei
- Xinyu Wen
- Erica Pehrsson

... ...

ABCS NCI Frederick

- Vishal Koparde
- Anney Che

... ...

- Parthav Jailwala
- Uma Mudunuri

CCDI, CBIIT and AWS

- Subhashini Jagu
- Srujan Boppana
- Krish Seshadri
- Sue Pan
- Kevin Sayers

ODSS support of NCI data science project for exploratory use of the cloud via the STRIDES Initiative in FY2023, 2024

